Estudos de Economia, vol. III, n.º 2, Jan.-Mar., 1983

A DISTRIBUIÇÃO LOGNORMAL: SUA APLICAÇÃO A DADOS PORTUGUESES DE DISTRIBUIÇÃO DO RENDIMENTO

Jorge Santos (*)

1 — Introdução

É frequente a análise dos dados sobre repartição dos rendimentos em termos de inúmeras medidas de estatística descritiva de que são exemplo os intervalos inter-quantis, o coeficiente de Gini, ou o coeficiente de variação. O objectivo é quase sempre o de quantificar a desigualdade existente, ainda que o conceito de igualdade e o grau de desigualdade «aceitável» sejam polémicos e quase nunca abordados.

Um passo em frente no sentido de uma quantificação «normativa» da desigualdade é o índice de Atkinson (1970), discutíveis que possam ser algumas das suas hipóteses. Mas o que é comum a quase todos esses estudos é o facto de não se admitir qualquer distribuição teórica subjacente aos dados a analisar.

Ora acontece que uma distribuição estatística teórica apareceu precisamente da análise desses dados — a distribuição ou «lei» de Pareto; e que considerável esforço tem sido despendido pelos economistas a testar essa e outras distribuições, bem como a compreender quais os mecanismos económicos que a elas poderiam conduzir.

À euforia inicial de Pareto, que descobriu um ajustamento quase perfeito da função

$$N = AX^{-\alpha}$$

N — número de rendimentos acima de um certo rendimento X
A e α — constantes.

Aos dados publicados na época para diferentes períodos e diferentes países sobreveio uma certa desilusão: o desenvolvimento do aparelho estatístico veio a demonstrar que os dados então existentes, e que se referiam aos rendimentos sujeitos a impostos (rendimentos mais altos), não tinham a mesma distribuição dos rendimentos mais baixos (1).

(*) Assistente do ISE, bolseiro do INIC. O presente artigo é baseado num capítulo da tese de Master, submetida à Universidade de Bristol em Outubro de 1980 e intitulada *Inequality Measures — The Lognormal Distribution and the Portuguese Experience*.

(1) Apesar de tudo ainda hoje se considera que a distribuição de Pareto é uma descrição bastante razoável dos 20 % de mais altos rendimentos. Esta é a razão pela qual esta distribuição é vulgarmente utilizada em interpolações realizadas nas classes de rendimentos superiores, nomeadamente na estimativa do ponto médio da classe aberta.

A distribuição «típica» dos rendimentos constatou-se ser a seguinte (fig. 1):

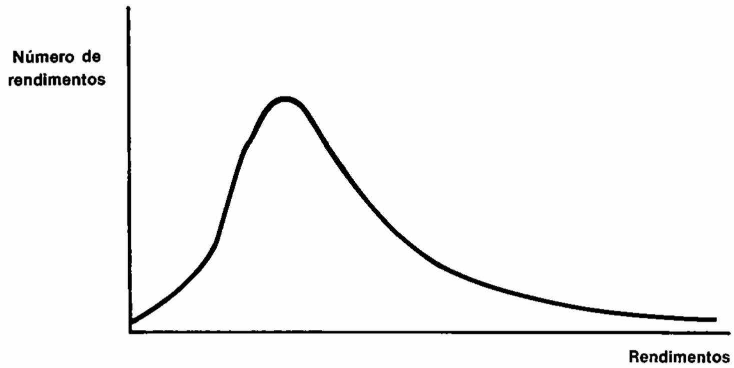

FIGURA 1

Distribuição típica dos rendimentos

isto é, uma curva unimodal, assimétrica e desviada à direita.

De todas as distribuições teóricas, a que melhor se aproxima da distribuição empírica parece ser a distribuição lognormal. Não admira que se tenham realizado várias tentativas para estimar os seus parâmetros em vários países e em várias épocas, com maior ou menor grau de sucesso. A conclusão geral de todas essas experiências tem sido a de que quanto mais homogéneos são os que recebem o rendimento tanto melhor será o ajustamento. Embora a lognormal seja uma distribuição que dá uma razoável aproximação à globalidade dos rendimentos, tem-se revelado bastante pobre no que se refere aos rendimentos mais altos, onde continua a prevalecer a distribuição de Pareto.

No nosso caso, e para Portugal, fez-se uma tentativa de ajustamento da lognormal às distribuições empíricas dos salários de base dos homens e mulheres trabalhando na indústria, com mais de 20 anos de idade, vivendo no continente e regiões autónomas (excepto para 1980 — só continente), e referentes ao ano de 1972 e ao período de 1975 a 1980. A fonte dos dados reside nos inquéritos por amostragem realizados pelos Serviços de Estatística do Ministério do Trabalho e intitulados «Níveis de Remuneração».

As vantagens de um possível bom ajustamento seriam:

Permitir a descrição de toda a distribuição a partir de muitos poucos parâmetros;

A distribuição lognormal ser relativamente simples de usar, com propriedades matemáticas bem conhecidas e os seus parâmetros poderem ser facilmente ligados a importantes medidas económicas, como se verá;

Para construir modelos econométricos de distribuição de rendimentos [2], o conhecimento de que existe uma distribuição teórica que fornece um bom ajustamento à distribuição empírica, se não essencial, facilitaria pelo menos essa tarefa.

(2) V., por exemplo, Metcalf (1972).

Há a salientar no entanto que o presente estudo é limitado a uma classe restrita de rendimentos e de pessoas. Contudo, a série já relativamente longa de dados existentes, bem como a sua confiança, fazem com que este seja um começo natural.

As secções seguintes serão dedicadas à breve descrição matemática da lognormal, aos argumentos apresentados justificativos da sua génese, à estimação dos seus parâmetros e finalmente à apresentação dos resultados obtidos para Portugal.

2 — **Breve descrição matemática da Lognormal**

Pode-se considerar a distribuição lognormal com dois, três ou mesmo quatro parâmetros. Contudo, para os nossos propósitos, só os dois primeiros casos nos interessam. O caso mais simples, que combina algumas das propriedades da distribuição normal com as da função logarítmica, é, obviamente, a lognormal com dois parâmetros, que pode ser matematicamente descrita como segue:

Lognormal com dois parâmetros:

Seja X uma variável positiva $(0 < x < \infty)$ tal que $Y = \ln X$ é normalmente distribuída com média μ e variância σ^2. Dizemos então que X é uma variável Λ. Podemos escrever

$$X \text{ é } \Lambda\left(\mu,\sigma^2\right) \text{ e}$$
$$Y \text{ é } N\left(\mu,\sigma^2\right)$$

Sejam $\Lambda\,(x \mid \mu,\sigma^2)$ e $N\,(y \mid \mu,\sigma^2)$ as funções de distribuição de X e de Y, respectivamente, isto é,

$$\Lambda\left(x \mid \mu,\sigma^2\right) = P\left[X \leq x\right],$$
$$N\left(y \mid \mu,\sigma^2\right) = P\left[Y \leq y\right];$$

então

$$\Lambda\,(x) = N\,(\ln x) \qquad x > 0$$

pelo que

$$\Lambda\,(x) = 0 \qquad x \leq 0$$

A função de densidade é deduzida imediatamente da correspondente para a distribuição normal, e é dada pela expressão seguinte:

$$d\,\Lambda\,(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right] \qquad dx \qquad (x > 0)$$

A partir daqui facilmente se chegará aos seguintes resultados:

A moda (única) é o valor de X tal que

$$x = \exp(\mu - \sigma^2)$$

A média α e a variância β^2 de X são dadas por

$$\begin{aligned} \alpha &= \exp(\mu + 0.5\sigma^2) \\ \beta^2 &= \exp(2\mu + \sigma^2)(\exp \sigma^2 - 1) \\ &= \alpha^2\eta^2 \end{aligned}$$

onde $\eta^2 = \exp \sigma^2 - 1$. Como $\eta^2 = \beta^2 / \alpha^2$, η é o coeficiente de variação da distribuição.

As expressões para os quantis (particularmente úteis no processo de estimação, como se verá) podem ser expressas por

$$\xi_q = \exp(\mu + v_q\sigma)$$

onde ξ_q e v_q *são os quantis de ordem* q para $\Lambda(\mu,\sigma^2)$ e $N(0,1)$, respectivamente. Isto implica, por exemplo, que o 1.° e o 3.° quartis sejam $x = \exp(\mu - 0.67\sigma)$ e $x = \exp(\mu + 0.67\sigma)$, respectivamente, e que a mediana seja em $x = \exp(\mu)$.

Uma relação muito importante liga os parâmetros da lognormal com uma não menos importante medida de desigualdade — o coeficiente de Gini (entendido como o valor do ratio da área entre a curva de Lorenz e a linha de igualdade absoluta, e a área para baixo dessa linha). Assim

$$G = 2N\left(\frac{\sigma}{\sqrt{2}} \mid 0,1\right) - 1$$

onde G é o coeficiente de Gini.

Uma importante conclusão se retira da expressão acima: σ (o desvio padrão da distribuição lognormal) é ele próprio uma medida de desigualdade. Pode-se também provar que as curvas de Lorenz correspondentes a distribuições lognormais de dois parâmetros não se cruzam. Os resultados de Atkinson (1970) permitem então concluir que será indiferente escolher σ ou qualquer outra medida de desigualdade, pois as distribuições serão ordenadas da mesma maneira, sem qualquer ambiguidade. Prova-se igualmente que as curvas de Lorenz são simétricas.

Lognormal com três parâmetros:

Quando $X - \tau$ é $\Lambda(\mu,\sigma^2)$ e não X, dizemos que X é uma lognormal com três parâmetros, e escrevemos:

$$X \text{ é } \Lambda(\tau,\mu,\sigma^2)$$

O domínio de X é $\tau < x < \infty$, e, como τ define o limite inferior de X, pode ser chamado o limiar da distribuição, embora seja também chamado o parâmetro de deslocamento. A função de distribuição é dada por

$$\Lambda(x \mid \tau,\mu,\sigma^2) = 0 \qquad (x \leq \tau)$$
$$\Lambda(x \mid \tau,\mu,\sigma^2) = \Lambda(x - \tau \mid \mu,\sigma^2) \qquad (x > \tau)$$

A função de densidade é também a de $\Lambda(\mu,\sigma^2)$ deslocada de τ. Todas as medidas de localização são pois deslocadas de τ: a média é no ponto $x = \tau + \alpha$, a mediana em $x = \tau + exp(\mu)$ e a moda em $x = \tau + exp(\mu - \sigma^2)$. Os quantis são dados por $\xi_q + \tau$.

Para terminar, refira-se que, em relação às medidas de desigualdade, o coeficiente de variação é dado por

$$\eta' = \frac{\eta}{1 + \tau / \alpha}$$

e o coeficiente de Gini por

$$G' = \frac{\alpha}{\tau + \alpha}\left[2N\left(\frac{\alpha}{\sqrt{2}} \,\Big|\, 0,1\right) - 1\right]$$

e que as curvas de Lorenz se podem intersectar.

3 — A génese da distribuição lognormal

A ideia geral do teorema do limite central é a de que se $Y = \sum_{i=1}^{n} X_i$, em em que as X_i são variáveis aleatórias independentes, com a mesma distribuição e variância finita, então a distribuição de Y tenderá para a distribuição normal quando $n \longrightarrow \infty$. Poder-se-á então concluir que se $Y = \prod_{i=1}^{n} X_i$, a distribuição de Y será assimptoticamente lognormal. Isto significa que, se tencionarmos explicar a lognormalidade do rendimento, devemos indicar quais os factores que agem multiplicativamente para gerar os rendimentos observados.

Existem duas teorias principais: uma diz que o rendimento é o resultado de um processo aleatório infinito que age multiplicativamente, e a outra que o rendimento é o produto de factores independentes operando simultaneamente num ponto do tempo.

Consideremos a primeira teoria:

Suponha-se uma variável X sujeita à «lei do efeito proporcional» — o acréscimo na variável em qualquer passo de um processo de variação é uma proporção aleatória do valor prévio da variável.

O processo de Markov é então:

$$X_t - X_{t-1} = \epsilon_t X_{t-1}$$

ou

$$\epsilon_t = \frac{X_t - X_{t-1}}{X_{t-1}}$$

então,

$$\sum_{t=1}^{n} \epsilon_t = \sum_{t=1}^{n} \frac{X_t - X_{t-1}}{X_{t-1}}$$

Mas

$$\sum_{t=1}^{n} \frac{X_t - X_{t-1}}{X_{t-1}} \approx \int_{X_\emptyset}^{X_n} \frac{dX}{X} = lnX_n - lnX_\emptyset$$

pelo que

$$lnX_n \approx lnX_\emptyset + \epsilon_1 + \epsilon_2 + \ldots + \epsilon_n$$

podendo-se então concluir pela forma aditiva do teorema do limite central que lnX_n é assimptoticamente normalmente distribuída com dois parâmetros.

A distribuição com três parâmetros seria gerada pelo processo:

$$X_t - X_{t-1} = \epsilon_t (X_{t-1} - \tau)$$

Entre outros, Champernowne (1953) e Kalecki (1945) utilizaram estas ideias básicas para criarem modelos que implicariam uma distribuição de rendimentos lognormal. No entanto, algumas das suas hipóteses são bastante criticáveis na sua comparação com a realidade. Refira-se que num modelo do processo aleatório X_n, concebido como uma sequência de acontecimentos no tempo, quanto maior é o número de períodos tanto maior será a variância de X_n. Quando referido à distribuição de rendimentos, isso significaria crescente desigualdade, o que é contrário à evidência empírica. Para evitá-lo, tanto Kalecki como Champernowne introduziram algumas restrições puramente matemáticas. Rutherford (1955) tentou resolver este problema de uma maneira diferente: salientou que embora a variância do rendimento tenda a aumentar com a idade (pelo menos em algumas ocupações), ela tenderia a ser compensada pela entrada de novos receptores de rendimento e pela morte dos mais velhos. Se se considerar a população dividida em coortes, é fácil pensar em variância crescente dentro de cada coorte e constância da variância para a população como um todo.

Se aplicarmos a lei do efeito proporcional não a um processo no tempo mas a um grande número de factores independentes que se combinam multiplicativamente num dado momento, estamos dentro do segundo tipo de teorias que tentam explicar a distribuição lognormal. Suponha-se, como Roy (1950) o faz, que o produto de um trabalhador é o resultado de várias «habilidades», como sejam a velocidade, idade, educação, saúde, etc. Se estas «habilidades» forem normalmente distribuídas, e sendo cada uma uma variável aleatória, então a hipótese crucial é a de que elas se combinam não aditivamente mas multiplicativamente. A distribuição do produto tenderia então para a lognormalidade. E, como Roy salienta, «no que se refere a simplicidade não há razão para colocar a multiplicação mais baixo na escala do que a adição». Contudo, o que é uma «habilidade», e porque deveria ser ela normalmente distribuída? Embora haja uma maior ou menor ampla evidência que as características físicas dos indivíduos (p. ex. altura) são normalmente distribuídas, nada há que prove que sucede o mesmo com as características mentais ou psicológicas.

Poder-se-á então concluir que não há uma teoria convincente para a geração da lognormal no que se refere à distribuição dos rendimentos: é conhecido o processo aleatório básico subjacente a tal distribuição (a lei do efeito proporcional), e que os factores (variáveis) devem agir multiplicativamente em vez de aditivamente. Contudo, os mecanismos económicos que podem eventualmente conduzir a tal distribuição são ainda muito obscuros. A exagerada ênfase dada aos processos aleatórios ou a modelos matemáticos que possam conduzir à distribuição lognormal pode dar algumas vezes como resultado que a distribuição de rendimentos não seja considerada e estudada como um fenómeno global económico e sociológico. A estratificação da sociedade em classes, a relativa imobilidade entre elas, a influência da riqueza inicial e herdada, a educação, o ambiente familiar, etc., poderão ser quase completamente esquecidos.

O uso da distribuição lognormal na distribuição dos rendimentos continua a ter assim um carácter experimental com uma base teórica insuficiente. E foi como tal que ele foi considerado neste estudo de aplicação para Portugal.

4 — **Métodos de estimação**

No sentido de tentar estimar os parâmetros da lognormal vários métodos são possíveis; Aitchison e Brown (1957) realizaram o que hoje se pode chamar um estudo de Monte Carlo no sentido de tentar determinar qual o melhor método. As alternativas eram: o método dos quantis, o método dos momentos, o método gráfico (usando papel logaritmo-probabilístico), métodos mistos (usando simultaneamente alguns dos métodos atrás citados) e o método da máxima verosimilhança. Este último era considerado o método ideal, mas os meios de cálculo então existentes constituíam um poderoso obstáculo à sua utilização. Dos restantes métodos concluiram da preferência pelo método dos quantis.

No presente estudo faz-se uma comparação dos resultados obtidos com o método dos quantis e o método da máxima verosimilhança. Antes de descrever brevemente estes dois métodos dever-se-á referir que se optou por estimar os parâmetros tanto da lognormal com dois parâmetros como da com três parâmetros. A razão foi a de que, pelo exame das curvas de Lorenz (previamente desenhadas) se verificou serem elas razoavelmente simétricas e com poucos, embora alguns, problemas de cruzamento (nomeadamente para os valores mais altos da distribuição de 1972 em relação a todas as outras e na comparação dos anos de 1976 e 1980). Não havia pois qualquer razão válida para excluir *a priori* qualquer dos casos. No entanto, a existência desses (poucos) casos de cruzamento, bem como a evolução do salário mínimo em Portugal sugeririam um melhor ajustamento pela distribuição com três parâmetros, o que aliás não se verificou, como se verá.

Passemos agora a descrever brevemente os dois métodos de estimação utilizados (4):

Máxima Verosimilhança:

Suponha-se dada uma amostra S_n de tamanho n de $\Lambda(\mu, \sigma^2)$ consistindo nas observações $x_1, x_2, \ldots, x_n$. O problema é encontrar funções de amostra que sejam bons estimadores de μ e σ^2. As estimativas da máxima verosimilhança são os valores dos parâmetros da distribuição que «mais verosimilmente» gerou a amostra observada. A função de verosimilhança da amostra é dada por

$$L = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \prod_{i=1}^{n} x_i} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(\ln x_i - \mu)^2\right]$$

Calculando os valores de μ e σ^2, que maximizam a função de verosimilhança (ou o seu logaritmo), encontramos os estimadores de máxima verosimilhança m e s^2 de μ e σ^2, respectivamente. Esses estimadores são dados pelas expressões

$$m = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i - m)^2$$

isto é, m e s^2 são a média e a variância da amostra transformada:

$$s^2 = \frac{n-1}{n} v_y^2 \quad \text{em que} \quad v_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \; y_i = \ln x_i$$

O estimador s^2 é pois enviezado mas consistente. Se usarmos como nosso estimador da variância $s^2 = v_y^2$, então m e s^2 são estimadores não enviezados de variância mínima de μ e σ^2. Da teoria normal podemos facilmente deduzir que as variâncias de m e s^2 são dadas por

$$V(m) = \frac{\sigma^2}{n}$$

$$V(s^2) = \frac{2\sigma^4}{n-1} \approx \frac{2\sigma^4}{n}$$

(4) Expor-se-á no método de máxima verosimilhança unicamente o caso de observações não agrupadas em classes. Para informações mais detalhadas, ver Aitchison e Brown (1957).

No caso da lognormal com três parâmetros, a função de verosimilhança é

$$L = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \prod_{i=1}^{n} (x_i - \tau)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(ln(x_i - \tau) - \mu)^2\right]$$

Como referido em Cohen e Whitten (1980), Hill (1963) demonstrou a existência de caminhos ao longo dos quais a função de verosimilhança de qualquer amostra ordenada, $x_1, \ldots, x_n$ tende para ∞ à medida que (τ, μ, σ^2) se aproximam de $(x_1, -\infty, \infty)$. Isto significa que o máximo global conduz às estimativas $t = x_1$, $m = -\infty$ e $s^2 = \infty$ $_{(em\ que}$ t, m e s^2 são as estimativas de τ, μ, σ^2, respectivamente). Se usarmos o método habitual de igualar as derivadas parciais da função log de verosimilhança a zero e resolver então o resultante sistema de equações simultâneas, a solução será um máximo local. Estas soluções têm sido encontradas razoáveis se comparadas com outros métodos de estimação.

As equações estimadoras locais de máxima verosimilhança são então,

$$\frac{\delta lnL}{\delta\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}[ln\ (x_i - \tau) - \mu] = 0$$

$$\frac{\delta lnL}{\delta\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^{n}[ln\ (x_i - \tau) - \mu]^2 = 0$$

$$\frac{\delta lnL}{\delta\tau} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{[ln\ (x_i - \tau) - \mu]}{x_i - \tau} + \sum_{i=1}^{n}(x_i - \tau)^{-1} = 0$$

Eliminando σ^2 e μ, a equação em τ é dada por

$$\lambda(\tau) = \sum_{i=1}^{n}(x_i - \tau)^{-1}\left\{\sum_{i=1}^{n} ln\ (x_i - \tau) - \sum_{i=1}^{n} ln^2\ (x_i - \tau) + \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n} ln\ (x_i - \tau)\right]^2\right\} -$$

$$- n\sum_{i=1}^{n}\frac{ln\ (x_i - \tau)}{(x_i - \tau)} = 0$$

Esta equação é então resolvida iterativamente para a estimativa local de máxima verosimilhança de τ, t. Os outros estimadores são dados por,

$$m = \left(\frac{1}{n}\right)\sum_{i=1}^{n} ln\ (x_i - t)$$

$$s^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n} ln^2\ (x_i - t) - \left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} ln^2\ (x_i - t)\right]^2$$

Método dos quantis:

No caso da lognormal com dois parâmetros vimos já que a relação entre os quantis de ordem q de Λ (μ, σ^2)—ξ_q, e de N (0, 1)—v_q *é dada por*

$$\xi_q = \exp\ (\mu + v_q\sigma)$$

6

Se denotarmos os quantis da amostra de ordem q_1 e q_2 ($q_1<q_2$) por x_{q1} e x_{q2} os estimadores quantis de μ e σ, respectivamente m' e s', são dados pelas expressões

$$x_{q1} = \exp(m' + v_{q1} s')$$

$$x_{q2} = \exp(m' + v_{q2} s')$$

A máxima eficiência é obtida quando os quantis são simétricos. Como a distribuição normal é simétrica, então

$$v_{1-q} = -v_q = v \text{ onde } q<\frac{1}{2}$$

pelo que as expressões para m' e s' podem ser escritas como sendo

$$m' = \frac{1}{2}(ln x_{1-q} + ln x_q)$$

$$s' = \frac{1}{2v}(ln x_{1-q} + ln x_q)$$

No caso da distribuição com três parâmetros, teremos de utilizar três quantis. Aitchison e Brown sugeriram o uso da mediana e dos quantis com $q=0.05$, 0.10 ou 0.20. As equações são agora

$$x_q = t + \exp(m' - vs)$$

$$x_{\frac{1}{2}} = t + \exp m'$$

$$x_{1-q} = t + \exp(m' + vs)$$

pelo que as expressões para os estimadores são dadas por

$$s = \frac{1}{v}[ln(x_{1-q} - x_{\frac{1}{2}}) - ln(x_{\frac{1}{2}} - x_q)]$$

$$m' = ln(x_{\frac{1}{2}} - x_q) - ln(1 - e^{-vs})$$

$$t' = x_{\frac{1}{2}} - e^{m'}$$

5 — Resultados para Portugal

As estimativas para os parâmetros da lognormal com 2 e 3 parâmetros, usando o método de máxima verosimilhança (5), são apresentados nos quadros 1 e 2. Os valores entre parênteses indicam os erros padrões.

---

(5) Na sua computação foi usado o programa MLP, versão 3.04-1976, *Lawes Agricultural Trust, Rothamsted Experimental Station.*

Quadro 1

Método de máxima verosimilhança

Lognormal de 2 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Média	.919.32 (.00102)	1.72297 (.00084)	1.89707 (.00069)	1.9952 (.00065)	2.12852 (.00053)	2.28900 (.00056)	2.44144 (.00055)
Desvio padrão	.51276 (.00078)	.38661 (.00060)	.32061 (.00050)	.32082 (.00047)	.29215 (.00038)	.31622 (.00041)	.31084 (.00040)

Quadro 2

Método de máxima verosimilhança

Lognormal de 3 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Parâmetro de deslocamento	.10711 (.00554)	1.76458 (.00525)	2.51184 (.00886)	2.95570 (.01223)	2.84996 (.00894)	3.10399 (.01710)	4.45487 (.00844)
Média	.87033 (.00281)	1.29584 (.00208)	1.37568 (.00272)	1.42511 (.00358)	1.68237 (.00195)	1.87839 (.00300)	1.90394 (.00165)
Desvio padrão	.54075 (.00177)	.56786 (.00130)	.51956 (.00151)	.54464 (.00197)	.44661 (.00098)	.46706 (.00146)	.50799 (.00096)

Como os erros padrões são muito pequenos, todos os valores estimados são estatisticamente significativos.

Comparando as distribuições geradas por estes parâmetros com as distribuições empíricas, e usando o teste de Kolmogorov-Smirnov, concluiu-se não se poder rejeitar a hipótese de as distribuições da amostra provirem de distribuições lognormais com os parâmetros estimados.

Contudo, Massey (1951) sugeriu que quando os parâmetros são estimados a partir de dados de amostragem este teste é «conservador», no sentido de que a probabilidade de erro de tipo I é menor do que a dada na maior parte das tabelas, pelo que deverão pôr-se reservas quanto a esta conclusão.

Os maiores desvios entre as distribuições estimadas e empíricas situaram-se na zona das duas modas das frequências empíricas (não se esqueça que se analisou a distribuição conjunta de homens e mulheres). As caudas das distribuições também foram em geral pobremente estimadas. A figura 2 dá uma ideia da relação entre as distribuições empíricas e estimadas.

No sentido de se obter melhores estimativas ainda se iniciou um processo iterativo em que em cada iteração se usaram os valores obtidos na iteração prévia para as últimas seis classes. Isto conduzia a valores mais próximos da distribuição empírica, mas o processo, por muito custoso e consumidor de tempo, foi abandonado. Saliente-se que neste estudo nunca se procurou ajustar a distribuição de Pareto às classes de mais altos rendimentos. No que se refere às duas modas, estimaram-se também os parâmetros para a lognormal usando para todos os anos as distribuições empíricas correspondentes só aos homens, mas as melhorias não foram significativas (6).

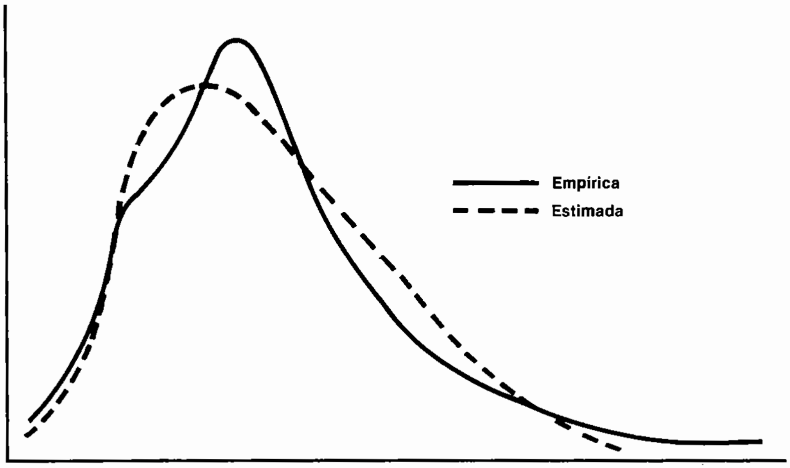

FIGURA 2

Distribuição empírica e estimada

Os valores para os índices de Gini e coeficiente de variação implícitos nos parâmetros estimados estão indicados nos quadros 3 e 4.

Quadro 3

Coeficientes de Gini e coeficientes de variação
Método de máxima verosimilhança

Lognormal de 2 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Coeficiente de Gini	.283	.215	.179	.180	.164	.177	.174
Coeficiente de variação......	.548	.402	.329	.329(3)	.299	.324	.319

Quadro 4

Coeficientes de Gini e coeficientes de variação
Método de máxima verosimilhança

Lognormal de 3 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Coeficiente de Gini	.287	.242	.211	.217	.187	.199	.205
Coeficiente de variação......	.561	.478	.410	.426	.355	.381	.397

(6) V. Santos (1980).

Em termos de desigualdade e no caso da lognormal de 2 parâmetros, verificar-se-ia diminuição até 1976, um ligeiro aumento em 1977 (ou uma certa indeterminação se levarmos em linha de conta os intervalos de confiança das estimativas), continuação da tendência decrescente em 1978 e finalmente uma brusca subida em 1979, sustentada em 1980. Recorde-se que estas conclusões poderiam ter sido retiradas imediatamente da análise dos valores dos desvios padrões estimados. No caso dos três parâmetros, a tendência é praticamente a mesma, embora já sem dúvidas quanto a 1977. Saliente-se aqui que os resultados obtidos directamente das distribuições empíricas nos davam uma diminuição da desigualdade até 1978, com um agravamento em 1979, continuado em 1980, embora com valores para o índice de Gini inferiores.

Quanto à estimação utilizando o método dos quantis, foram utilizados 3 pares de quantis simétricos: (0.20, 0.80), (0.10, 0.90) e (0.05, 0.95). A mediana foi o terceiro quantil a ser utilizado na estimação da lognormal de 3 parâmetros.

Uma razão poderia levar-nos a usar só os quantis (0.20, 0.80): evitar interpolações na classe aberta. Como esse problema só surge em 1972, decidiu-se utilizar para este ano só os quantis 0.20 e 0.80 e realizar a estimação para os outros anos usando os três pares atrás referidos. Os valores dos quantis utilizados estão indicados no quadro 5.

Quadro 5

Valores dos quantis

(Unidades = milhares de escudos)

	1972	1975	1976	1977	1978	1979	1980
.05	-	3.202	4.160	4.429	5.275	6.031	7.342
.10	-	3.437	4.418	4.905	5.765	6.552	7.781
.20	1.632	3.906	4.933	5.634	6.628	7.582	8.926
.50 (median)	2.491	5.490	6.634	7.232	8.379	9.897	11.076
.80	3.833	7.473	8.368	9.132	10.059	12.061	14.128
.90	-	9.035	9.816	11.0	11.947	14.452	17.111
.95	-	11.154	11.667	13.167	14.455	17.1	21.5

Os resultados foram muito semelhantes aos obtidos com a estimação de máxima verosimilhança no que se refere à lognormal com 2 parâmetros. No que respeita à lognormal com 3 parâmetros, a variação nos resultados foi consideravelmente maior e o facto de maior saliência foi a obtenção de valores negativos na estimação do desvio padrão (em 1978 e 1979) quando utilizados os quantis 0.20 e 0.80. Saliente-se no entanto que no presente caso, e em termos de aproximação aos dados empíricos, a lognormal com 2 parâmetros não é inferior à de 3 parâmetros, contrariamente ao que se poderia pensar *a priori* e como já mencionado.

A título de ilustração [v. Santos (1980) para uma completa listagem dos resultados] apresentam-se a seguir os valores estimados para a distribuição com 2 parâmetros utilizando os quantis (0.20, 0.80) e com 3 parâmetros (quantis 0.10 e 0.90), bem como as medidas de desigualdade implícitas para este último caso:

Quadro 6

Método dos quantis — Quantis 0.20 e 0.80

Lognormal de 2 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Média....................	.917	1.687	1.860	1.970	2.100	2.258	2.419
Desvio padrão	.507	.385	.314	.287	.248	.276	.273

Quadro 7

Método dos quantis — Quantis 0.10 e 0.90

Lognormal de 3 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Parâmetros de deslocamento	-	.610	— .666	1.150	1.960	—2.706	3.815
Média....................	-	1.585	1.988	1.805	1.859	2.534	1.983
Desvio padrão	-	.426	.282	.376	.243	.241	.472

Quadro 8

Coeficientes de Gini e coeficientes de variação
Método dos quantis — Quantis 0.10 e 0.90

Lognormal de 3 parâmetros

	1972	1975	1976	1977	1978	1979	1980
Coeficiente de Gini	-	.215	.175	.184	.110	.184	.198
Coeficiente de variação......	-	.405	.319	.342	.200	.332	.379

Em anexo apresentam-se para os anos de 1979 e 1980 quadros com as distribuições empíricas e as distribuições resultantes das estimações pelo método de máxima verosimilhança (lognormal com 2 e 3 parâmetros) e quantis (usando os quantis 0.20 e 0.80 e a lognormal de 2 parâmetros).

As conclusões gerais a retirar deste estudo empírico serão, pois:

A distribuição lognormal, quer com 2 quer com 3 parâmetros, é uma razoável aproximação aos dados empíricos estudados;

Contrariamente ao que se poderia pensar *a priori,* a lognormal com 3 parâmetros não fornece um melhor ajustamento em relação à de 2 parâmetros;

Os erros percentuais cumulativos são sempre relativamente pequenos e obedecem a um certo padrão. Contudo, dado um elevado número de observações, um pequeno erro percentual em cada classe pode representar um número significativo de indivíduos. Essa significância só poderá ser julgada pelos objectivos do estudo específico que estiver a ser realizado. Por outro lado, o conhecimento do grau de precisão possível obtido com o ajustamento da lognormal poderá ter um importante papel na determinação desses objectivos;

Quanto aos métodos de estimação utilizados, a conclusão geral que se pode retirar é a de que o método dos quantis, dada a sua simplicidade e resultados semelhantes, pode ser usado com sucesso em vez do método de máxima verosimilhança. Além disso fornece um importante grau de liberdade: o erro cumulativo é zero nos quantis escolhidos;

Refira-se, por último, que a continuação «lógica» deste estudo no sentido de uma melhoria dos ajustamentos conseguidos seria a utilização conjunta da lognormal e da distribuição de Pareto (esta última para os rendimentos mais altos). É uma linha de investigação em aberto.

ANEXO

QUADRO A1

1979

Unidade de rendimento – Milhares de escudos

Classe de rendimento	Frequência observada	Frequência estimada MMV — 2 parâmetros	Frequência estimada MMV — 3 parâmetros	Frequência estimada — 2 parâmetros Quantis (0.20, 0.80)
4- 5	0.4	1.6	0.4	0.9
5- 6	4.3	4.2	3.7	3.6
6- 7	9.6	8.1	9.3	8.3
7- 8	9.8	11.5	13.4	13.0
8- 9	11.9	13.2	14.4	15.4
9-10	15.6	13.1	13.3	15.1
10-11	18.1	11.8	11.2	13.0
11-12	10.0	9.8	8.8	10.1
12-13	4.9	7.6	6.8	7.2
13-14	4.0	5.8	5.0	4.9
14-15	3.1	4.2	3.7	3.2
15-16	1.9	3.0	2.7	2.0
16-17	1.3	2.1	2.0	1.3
17-18	1.0	1.4	1.4	0.8
18-19	0.7	1.0	1.0	0.5
19-20	0.5	0.6	0.8	0.3
20-25	1.5	1.1	1.6	0.4
25-30	0.7	0.1	0.4	0.0
> 30	0.7	0.0	0.1	0.0

QUADRO A2

1980

Unidade de rendimento – Milhares de escudos

Classe de rendimento	Frequência observada	Frequência estimada MMV — 2 parâmetros	Frequência estimada MMV — 3 parâmetros	Frequência estimada — 2 parâmetros Quantis (0.20, 0.80)
4.5- 5	0.1	0.4	0.0	0.1
5 - 6	0.3	1.5	0.2	0.9
6 - 6.5	0.3	1.5	0.8	1.2
6.5- 7	0.4	2.2	1.9	1.9
7 - 7.5	0.6	3.0	3.2	2.8
7.5- 8	10.8	3.7	4.5	3.7
8 - 9	8.1	9.4	11.7	10.2
9 -10	13.9	11.2	13.2	12.7
10 -11	14.5	11.7	12.7	13.4
11 -12	13.1	11.1	11.1	12.6
12 -13	10-5	9.9	9.2	10.8
13 -14	6.8	8.3	7.3	8.6
14 -15	4.7	6.7	5.7	6.5
15 -16	3.2	5.2	4.4	4.7
16 -17	2.5	4.0	3.4	3.3
17 -18	1.8	2.9	2.6	2.2
18 -19	1.6	2.2	2.0	1.5
19 -20	0.9	1.6	1.5	1.0
20 -25	3.0	3.1	3.5	1.6
25 -30	1.2	0.5	1.0	0.2
> 30	1.7	0.1	0.4	0.0

BIBLIOGRAFIA

AITCHISON, J. e BROWN, J. A. C. (1957) — *The Lognormal Distribution,* Cambridge University Press.

ATKINSON, A. B. (1970) — «On the Measurement of Inequality», *J. Econ. Theory.*

CHAMPERNOWNE, D. G. (1953) — «A Model of Income Distribution», *Econ. Journ.*

COHEN, A. C., e WHITTEN, B. J. (1980) — *Estimation in the Three-Parameter Lognormal Distribution,* J. of Amer. Statist Assoc.

KALECKI, M. (1945) — *On the Gibrat Distribution, Econometrica.*

MASSEY, F. J. (1951) — *The Kolmogorov-Smirnov Thest for Goodness-of-Fit,* J. of Amer. Statist. Assoc.

METCALF, C. E. (1972) — *An Econometric Model of the Income Distribution,* Markham Publishing, Chicago.

ROY, A. D. (1950) — «The distribution of earnings and of individual output», *Economic Journal.*

SANTOS, J. (1980) — *Inequality Measures — The Lognormal Distribution and the Portuguese Experience,* tese de mestrado submetida à Universidade de Bristol em Outubro de 1980.